Online Bayesian Nonparametrics for Group Detection

Matteo Zanotto
matteo.zanotto@iit.it
Loris Bazzani
loris.bazzani@iit.it
Marco Cristani
marco.cristani@iit.it
Vittorio Murino
vittorio.murino@iit.it

Pattern Analysis & Computer Vision
Istituto Italiano di Tecnologia
Via Morego 30 - 16163
Genova, Italy

Abstract

Group detection represents an emerging Computer Vision research topic motivated by the increasing interest in modelling the social behaviour of people. This paper presents an unsupervised method for group detection which is based on an online inference process over Dirichlet Process Mixture Models. Formally, groups are modelled as components of an infinite mixture and individuals are seen as observations generated from them. The proposed sequential variational framework allows to perform inference in real-time, while social constraints based on proxemics rules ensure the production of proper group hypotheses consistent with human perception. The results obtained on several datasets compare favourably with state-of-the-art approaches, setting the best performance in some of them.

1 Introduction

Social interactions are essential in our daily activities: people organise themselves in groups and share views, opinions, thoughts. Through the observations of social interactions many behavioural traits of the interlocutors can be inferred, as dominance or extroversion, or social characteristics of the interplay can be estimated [26]. For this reason, automatic modelling of interactional exchanges has become an active research topic in Computer Vision over the last few years, especially in video surveillance. The first step towards the analysis of social behaviour and the understanding of social interactions involves the identification of groups of people. Finding groups is very important to constrain the dynamics of people in tracking applications [20, 21, 28], in people recognition from still images [27], and in activity recognition tasks [9]. A group can be defined as several people who "interact on a regular basis, have affective ties with one another, share a common frame of reference, and are behavioural interdependent" [17]. For our (computer vision oriented) purposes, with group we mean a collection of (interacting) people which are spatially close, moving in the same direction or standing in a given local area.

This paper focuses on the automatic discovery of groups of people from videos in real surveillance scenarios. To this end, we employ a Dirichlet Process Mixture Model (DPMM)

[2] and propose an online inference algorithm to perform unsupervised learning. Specifically, groups are represented as components of a mixture model, and individuals are seen as observations generated from them. This model operates in a feature space defined by the output of a tracker (*i.e.*, the position and velocity of each individual at each time step) and embeds a "social" constraint driven by proxemics rules introduced by Hall [11]. Such constraint is based on the observation that the area shared by interactants is partitioned according to the so-called social distances, *i.e.* ranges of distances typically characterising social interactions. Such constraint allows to discard unlikely grouping configurations where the subjects are too far away from each other.

The proposed method has several desirable properties. First of all, it inherits all the advantages of the DPMMs, *i.e.* it deals with unlabelled data and it does not require prior knowledge about the number of groups (mixture components) that are to be found in a scene. Our method automatically adapts the number of groups to the observed data, coping with split, merge, initialization and removal phenomena, that characterise the high structural variability of groups. Another major advantage of our method is that it exploits an online inference strategy to update the parameters of the model. This is possible because group configurations evolve smoothly, so that the probabilistic model estimated at one time step can act as prior knowledge for the following one. This is accomplished by fitting the probabilistic model via sequential single-iteration variational inference in order to obtain the approximate posterior at each frame, which is subsequently used as a prior for the next frame (Fig. 1(b)). The employed Bayesian nonparametric approach demonstrated good flexibility and robustness, showing also real-time detection capabilities, up to about 42 fps (bounded by the performance of the people detection algorithm), using only position and velocity of people as a the feature representation.

In the next section the state of the art about group detection is outlined. Section 3 then presents the proposed approach. An extensive experimentation on public datasets is reported in Section 4, also providing critical comparative performance and conclusive remarks.

2 Related Work

One of the first papers showing interest in group modelling is the seminal work in computer animation by Reynolds [23]. He proposed a model that simulates the complex motion of birds (individuals) and flocks of birds (groups). Advanced techniques more specific for human motion [14] have been developed afterwards. One of the most used dynamical model is the Social Force Model (SFM) [12] that simulates the individuals in crowd as a gas-kinetic phenomenon, exploiting the concept of force from physics. Later, this model has been extended for dealing with groups [6]. Sochman and Hogg [25], on the other hand, propose a new agglomerative clustering method for group detection by inverting the SFM to infer its hidden parameters given tracking observations. Similarly, Ge *et al.* [10] use agglomerative clustering to group together tracking trajectories gathered in a time-window with fixed length. Pellegrini *et al.* [21] present a conditional random field to jointly estimate groups and the trajectory of individuals. Inference for this model is however too slow, making it unsuitable to applications where speed is critically important.

Yamaguchi *et al.* [28] propose an energy-based model to deal with group dynamics and prediction of destination in order to have a better estimate of the trajectories. Groups are defined as pairwise links among individuals. Pairwise features (*e.g.*, distance between two individuals) are fed into an SVM that predicts the group and non-group state of each pair of subjects. Jacques *et al.* [13] define a Voronoi diagram on the position of the individuals at

each frame, where the personal space is the area of the corresponding Voronoi polygon. A group is so heuristically defined as a set of adjacent (below a certain distance) polygons. The advantage of our approach with respect to [13, 28] is that it assumes that interactions depend on all members of the group not only on pairwise interactions.

Among the works addressing the analysis of group behaviour, it is interesting to highlight the differences between 2 classes of methods, *i.e.*, those considering only positional features at each time step [13, 21] (*e.g.*, the ground floor position), and those processing visual data over a time-window [3, 10, 25, 28] (*e.g.*, exploiting trajectories). The former provides the results at each time frame with no delay. They are usually faster because the single-frame output of standard detection/tracking algorithms is directly processed. The latter class of methods uses additional cues extracted from the images and/or trajectories, typically applied on top of detection/tracking results. Although considering a set of frames in a time window can make the chosen descriptors more robust, such information can often be noisy and unreliable, eventually affecting the performance of group detection, besides being more time consuming. For instance, in semi-stationary scenarios where people gather and talk, the head orientation is used [3, 8] to detect interactions. Cristani *et al.* [8] present a generalized Hough-voting strategy to estimate the F-formations [15], which are specific spatial configurations groups may assume. Bazzani *et al.* [3] introduce the inter-relation pattern matrix that condenses the pairwise relations between individuals, assuming the group relation to be transitive. The above techniques have two main drawbacks: the visual cues are not always available (because of low resolution, for example) and the assumption that head orientation gives the focus of attention of the person is not always valid. Moreover, estimating head orientation is still an issue in real, e.g., low resolution image, conditions.

Another distinction of the group detector is due to the nature of the learning method that is used: supervised or unsupervised. Supervised methods are characterised by the use of classifiers [28] and/or heuristics [3, 8, 12, 13] inherited from the application (*e.g.*, considering social psychology findings). The drawbacks of those techniques are the need of often demanding training phases and annotated data sets. Conversely, unsupervised methods analyse data directly in order to extract recurrent patterns exploiting the temporal redundancy of the group dynamics [4, 10]. Since almost no information is a-priori given, the latter approaches are more difficult to tackle, but the inherently high variability of the group detection problem and the lack of reliably annotated data sets on which to build our classifiers, make unsupervised methods the best choice for dealing with such kind of applications.

In the end, the proposed method has several novel characteristics which differentiates it from the state of the art. First of all, it only uses positional and velocity features at each time step, which are processed by an unsupervised online inference strategy. Such procedure also takes benefit from the use of proxemics cues, which validate and corroborate groups' hypotheses in a socially consistent way. Avoiding supervised learning strategies with their extensive training phases, using only low-level descriptors, and performing online inference, our method has unique features in the growing field of group behaviour analysis methods.

3 The Proposed Model

Modelling a set of groups can be seen as modelling a set of components of a mixture model living in an appropriate feature space. From this point of view, groups are seen as components and individuals are observations (or samples) drawn from them. Bayesian Nonparametrics, and specifically Dirichlet Process Mixture Models (DPMMs) [2] are particularly well suited for this task because they can represent mixture distributions with an unbounded

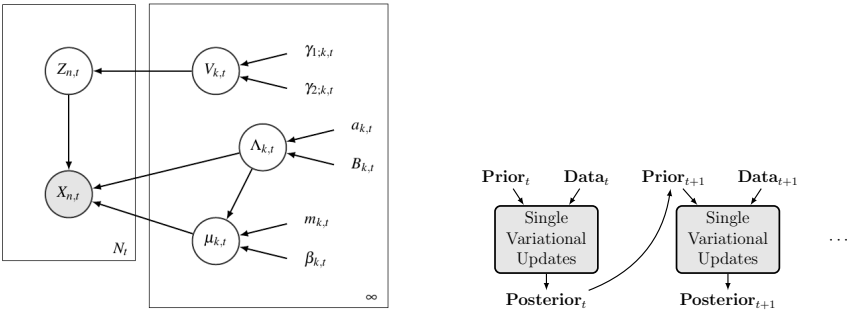


Figure 1: Graphical model representing the random variables and the parameters at time t.

number of components where the complexity of the model (i.e. the number of components) adapts to the observed data.

While the standard formulation of a DPMM considered in [5] handles observations that do not change over time, we propose a generalisation of their variational framework which circumvents some of the limitations of inference in Bayesian Nonparametric models. First of all our method allows to take into account the temporal evolution of the data without resorting to dynamical models [1] whose inference algorithms are affected by heavy computational loads. Secondly the proposed inference mechanism allows to produce results in real-time which is paramount for group detection in video sequences.

The proposed method relies on sequential variational inference that exploits the temporal correlation across consecutive frames to refine the detection of groups exploiting the evolution of the observations (Sec.3.1).

In addition, the framework is customised for dealing with people and groups, by taking into account proxemics notions. This bounds the maximal distance that separates two people in the same group, excluding grouping hypotheses with individuals which are too far away from one another (Sec.3.2).

3.1 Mathematical Formulation

Let us suppose that at time t we have N_t detected people; each subject is represented as a point $X_{n,t} = [x, y, \theta, \rho]$, where x and y represent the ground-floor position of the person, θ is the heading angle and ρ the velocity module. Each individual is interpreted as an observed sample coming from one of the infinitely many groups (component k).

Besides the parameters defining its distribution, each component has an associated probability mass depending on the parameters of the stick-breaking construction [24] used to model the Dirichlet Process prior [9]. A stick-breaking construction is a constructive process generating an infinite set of non-negative numbers summing to 1 by sequentially sampling from a series of *Beta* distributions. The obtained sequence can be interpreted as the mixing coefficients of a set of components defining a mixture model.

The graphical model associated to the described generative process is shown in Fig-

ure 1(a) where, at time step t, we have N_t points and

$$V_{k,t}|\gamma_{1;k,t},\gamma_{2;k,t} \sim Beta\left(\gamma_{1;k,t},\gamma_{2;k,t}\right) \tag{1}$$

$$Z_{n,t}|\{v_{1,t},v_{2,t},\ldots\} \sim Discrete\left(\pi\left(\mathbf{v}_t\right)\right) \tag{2}$$

$$\mathbf{\Lambda}_{k,t}|\mathbf{B}_{k,t},a_{k,t} \sim Wishart\left(\mathbf{B}_{k,t},a_{k,t}\right) \tag{3}$$

$$\mu_{k,t}|m_{k,t},\beta_{k,t},\mathbf{\Lambda}_{k,t} \sim Gaussian\left(m_{k,t},\left(\beta_{k,t}\mathbf{\Lambda}_{k,t}\right)^{-1}\right) \tag{4}$$

$$X_{n,t}|Z_{n,t} \sim Gaussian\left(\mu_{z_{n,t}},\mathbf{\Lambda}_{z_{n,t}}^{-1}\right) \tag{5}$$

where X_n represents the n^{th} data point, Z_n is an assignment variable relating each data point to the mixing components, V_k and the pair $(\mu_k,\mathbf{\Lambda}_k)$ represent the k^{th} mixture component in the stick-breaking construction [24] with $(\mu_k,\mathbf{\Lambda}_k)$ representing the location of the component in the parameter space and V_k defining the mixing proportions.

Among the parameters of the model (those represented without any circle in the graphical model of Figure 1(a)) γ_2 deserves special attention. The prior value associated to this parameter is α and is directly linked to the probability of generating new components (*i.e.* new groups) in the DPMM [5]. The effect of this free parameter on the final results will be analysed in Section 4.

One of the main requirements for video analysis algorithms is speed. Since people describe smooth trajectories when walking, the group configuration at each frame has strong correlation with that of the following frame and could be used as a prior belief for it. Starting from this consideration, a sequential variational inference framework has been derived to perform fast inference for group detection through DPMMs. This framework builds upon ideas by Neal and Hinton [19] and on the work by Blei and Jordan [5]. To introduce the least possible computational burden, single-iteration variational updates are performed on each frame and the obtained approximate posterior over the mixture model is used as a prior for the grouping configuration in the following frame. This is achieved by sequentially updating the parameters of the model (γ_1, γ_2, a, B, m, β) (see Figure 1(a)) estimated at time $t-1$ (prior for time t) using the data observed at time t as shown in Figure 1(b). Previous studies on the performance of online updates of parameters (such as online EM [18]) showed how these procedures can obtain good results, sometimes even better than the batch counterparts [18].

The sequential variational inference updates have been derived starting form the mean-field algorithm proposed by Blei and Jordan [5] and that reported by Penny [22].

As proposed by Blei and Jordan [5], mean field variational inference can be formulated using a family of variational distributions over $\theta = \{\mathbf{v},\mu,\mathbf{\Lambda},\mathbf{z}\}$ based on a truncated stick breaking construction with truncation level K

$$q(\theta) = \prod_{k=1}^{K-1} q_{\gamma_k}(v_k) \prod_{k=1}^{K} q_{\tau_k}(\mu_k,\mathbf{\Lambda}_k) \prod_{n=1}^{N} q_{\phi_n}(z_n) \tag{6}$$

In the formula above, n indexes the data points, k indexes the mixture components, $q_{\gamma_k} \sim Beta(\gamma_{1;k},\gamma_{2;k})$, $q_{\tau_k}(\mu_k,\Lambda_k)$ follows a Gaussian-Wishart model parametrised by $\tau_k = \{m_k,\beta_k,\mathbf{B}_k,a_k\}$ such that $q_{\tau_k}(\mu_k,\Lambda_k) \sim \mathcal{N}\left(\mu_k|m_k,(\beta_k\mathbf{\Lambda}_k)^{-1}\right)\mathcal{W}(\mathbf{\Lambda}_k|\mathbf{B}_k,a_k)$ and $q_{\phi_k}(z_n) \sim Discrete(\phi_n)$. The product over the $q_{\gamma_{;k}}$ stops at component $K-1$ since the last component absorbs all the residual probability mass of the stick-breaking construction and hence $q_{\gamma_K}(v_K) = 1$ [5].

Variational Bayes inference takes the form of an Expectation-Maximisation algorithm and can be divided in E-step and M-step. In the E-step the probability ϕ_n^k of each of the N points to belong to each of the K components is computed (see Supplementary Material for details).

Once all ϕ_n^k have been computed, the parameters of the distributions are updated in the M-step. After defining the following variables

$$\bar{N}_{k,t} = \sum_{n=1}^{N} \phi_n^{k,t} \qquad \bar{\mu}_{k,t} = \frac{1}{\bar{N}_{k,t}} \sum_{n=1}^{N} \phi_n^{k,t} \cdot x_{n,t} \tag{7a,b}$$

$$\bar{\Sigma}_{k,t} = \frac{1}{\bar{N}_{k,t}} \sum_{n=1}^{N} \phi_n^{k,t} \left(x_{n,t} - \bar{\mu}_{k,t}\right) \left(x_{n,t} - \bar{\mu}_{k,t}\right)' \tag{8}$$

the variational Bayes update formulas are used to update the parameters. The computed parameters define the posterior distributions at time-step t and will be used as prior for the following time-step from which the $t+1$ pedix is derived. In particular, the parameters $\gamma_{1;\cdot,t+1}$ and $\gamma_{2;\cdot,t+1}$ of the *Beta* distribution defining the mixing proportions are updated as

$$\gamma_{1;k,t+1} = \gamma_{1;k,t} + \bar{N}_{k,t} \qquad \gamma_{2;k,t+1} = \begin{cases} \gamma_{2;k,t} + \sum_{j=k+1}^{K} \bar{N}_{j,t} & \text{if } k < K \\ \alpha & \text{if } k = K \end{cases} \tag{9a,b}$$

with α being the scaling constant of the Dirichlet Process prior at time-step 0 [6]. The parameters of the distribution of the mean of each component of the mixture are updated as

$$m_{k,t+1} = \frac{\bar{N}_{k,t} \cdot \bar{\mu}_{k,t} + \beta_{k,t} \cdot m_{k,t}}{\bar{N}_{k,t} + \beta_{k,t}} \qquad \beta_{k,t+1} = \bar{N}_{k,t} + \beta_{k,t} \tag{10a,b}$$

Finally, the parameters of the distribution of the precision matrix of each component are updated according to

$$a_{k,t+1} = \bar{N}_{k,t} + a_{k,t} \qquad B_{k,t+1} = \bar{N}_{k,t} \cdot \bar{\Sigma}_{k,t} + \frac{\bar{N}_{k,t} \cdot \beta_{k,t} \left(\bar{\mu}_{k,t} - m_{k,t}\right) \left(\bar{\mu}_{k,t} - m_{k,t}\right)'}{\bar{N}_{k,t} + \beta_{k,t}} + B_{k,t} \tag{11a,b}$$

After the parameter initialisation, the presented variational EM procedure can be applied to the incoming data streams to perform inference over time.
To conclude the variational update, the components are sorted by decreasing number of assigned points.

The proposed method has two main advantages. First of all, inference is extremely fast and computational times are compatible with real time processing. Secondly, the spatial dynamics of the groups is not explicitly modelled, and this is especially valuable in scenarios where groups repeatedly split and merge.

After each update of the parameters, the model provides the probability each person has to belong to each of the groups. This depends on two different contributions. The first is the probability of the observation (person's position and velocity) under each Gaussian component which directly depends on the parameters m, β, B and a. The second is the probability of the Gaussian component itself which depends on the γ_1 and γ_2 parameters. People are assigned to groups performing hard assignment on the basis of the highest of these probabilities.

3.2 Social Constraint

Our clustering approach can be customized for managing individuals by considering elements of proxemics, which investigates how people use and organise the space they share with others. As is known from social psychology, people tend to unconsciously organise the space around them in concentric zones corresponding to different degrees of intimacy [11]. The idea is that the shorter the distance between two people, the higher the degree of intimacy. This analysis allows to define a limit distance, beyond which two individuals can be

	ETH	Hotel	Zara01	Zara02	Students003
People	360	390	148	204	434
Groups	74	59	45	58	109
2-people groups	50 (67.57%)	55 (93.22%)	36 (80.00%)	52 (89.66%)	88 (80.73%)
3-people groups	13 (17.57%)	3 (5.09%)	7 (15.56%)	6 (10.34%)	13 (11.93%)
≥4-people groups	11 (14.86%)	1 (1.69%)	2 (4.44%)	-	8 (7.34%)

Table 1: Statistics of the considered sequences.

considered not to be interacting with high probability. Recent studies define this distance as $r = 2$ meters [26], called social space. This notion helps us to refine the group hypotheses generated by our approach. When one of the detected groups does not fulfil the social constraint, *i.e.*, the members of the group are farther than r from each other, the corresponding mixture component is re-sampled. When this happens, all the parameters of such component (see Fig.1(a)) are re-initialised and the component is centred on the person which is less probable under the current mixture model by setting the mean to the position and speed of this person. This resampling procedure allows to propose new groups in areas badly approximated by the current mixture model.

4 Experiments and Discussion

Evaluating the performance of group detection algorithms is a hard task because of the lack of a commonly accepted evaluation protocol; for this reason, when comparing with other approaches, we adopt their metrics. In particular, we considered as competitors very recent state-of-the-art approaches [4, 28], focusing on heterogeneous benchmarks [16, 20]. In addition, we tested how the social constraint affects the performance of our approach.

Datasets. Two public benchmarks for group detection have been considered: the BIWI dataset [20], containing the `eth` and `hotel` sequences, and the Crowd by Example dataset [16], containing the `zara01`, `zara02`, `students003` sequences. Table 1 summarizes the statistics of each sequence as reported in the ground truth provided by Yamaguchi *et al.* [28]. Almost all the sequences present a high percentage of binary groups, while `eth` gives a wider coverage to groups of different size and represents a more balanced situation.

Evaluation Metrics. Our approach is compared with the pairwise model of [28], and the group detection and tracking model of [4]. In the comparison we used both the metric and the grouping ground truth presented in those works. In particular, we use the precision and recall defined over the pairwise relations as in [28], and the *1-False Positive rate* (1-FP), *1-False Negative rate* (1-FN) and *Group Detection Success Rate* (GDSR) averaged over each time step as in [4] (see the respective papers for the details on the metrics).

Results. We consider first the work of Yamaguchi *et al.* [28]; it is worth noting that their approach produces a unique relation matrix that, at the entry i, j, tells if subjects i and j have been detected as a group during the sequence. As consequence, their evaluation protocol aims at checking the similarity of the estimated relation matrix with the ground truth one. On the contrary, the proposed method gives a grouping configuration for each frame. In order to perform comparison, the results for each time step have to be merged in a single matrix. To this end, pairwise connections between two people are defined whenever they belonged to the same group for a certain fraction of of their tracks. Please note, such fraction is not a parameter of our model, and has merely been introduced to perform comparison with [28]. For this reason, Precision-Recall curves have been computed varying this parameter between 0 and 1. Figure 2 shows the curves obtained on the 5 sequences with and without introducing

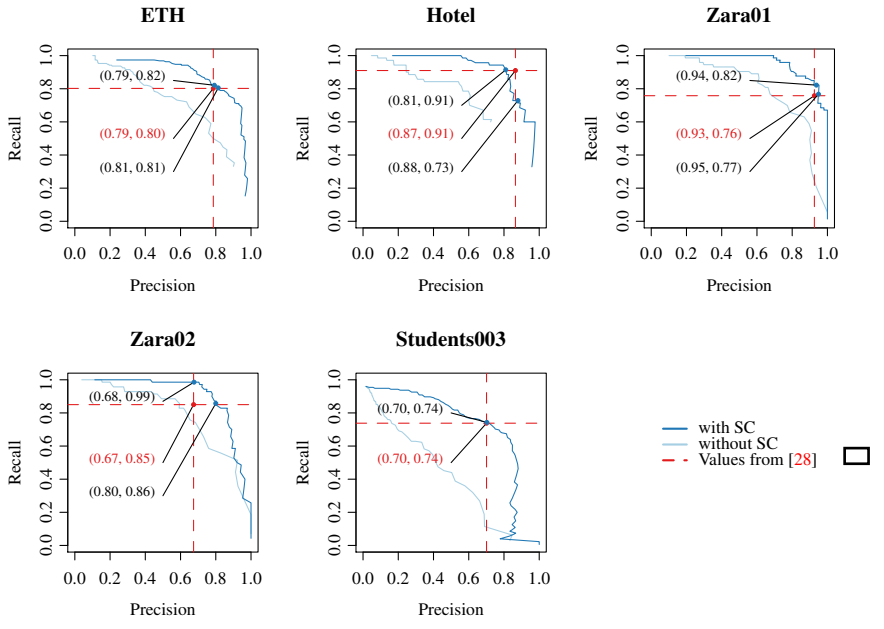


Figure 2: Precision-Recall curves with and without considering the social constraint + comparison with Yamaguchi *et al.* [28].

the social constraint (SC) in the model. The values for Precision and Recall reported in Table 3 of [28] (first column) are shown as dashed lines.

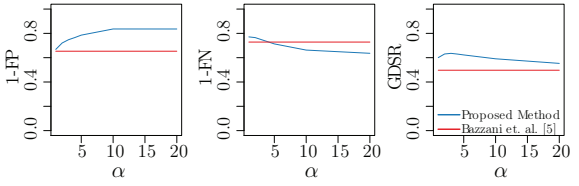
By visual inspection of Figure 2 some important facts can be observed. First of all, introducing the social constraint in the model leads to considerably better results. This proves the importance of including a mechanism guiding the inference over groups based on theories coming from social psychology.

When ignoring the social constraint, moreover, the point representing the results by Yamaguchi *et al.* [28] is always above the Precision-Recall curve by large margin.

After the introduction of the social constraint, on the other hand, such point is below the Precision-Recall curve obtained by the proposed method for all the sequences except for `hotel`. In order to better evaluate the extent of improvement, the results reported in [28] are annotated in Figure 2 along with the points of the Precision-Recall curve having closest Precision and closest Recall with respect to the benchmark method.

The worse performance on `hotel` is due to the fact that the sequence is extremely favorable for methods specifically designed to detect pairwise relations (rather than groups of any size) like [28]. The ground truth for this sequence, in fact, has a wide majority of couples of individuals (see Table 1)

In order to better assess the quality of the group detections, the method has been evaluated on the basis of the metrics proposed in [4] which explicitly consider the concept of group rather than decomposing it in a series of pairwise relations. As previously introduced, the `eth` sequence is the one which presents a more complete coverage of group sizes and hence is more appropriate to evaluate the performance of group detectors on diverse grouping configurations.



a)

	Bazzani et al. [5]	Online-BNP
1-FP	65.32	72.09
1-FN	72.84	76.52
GDSR	49.69	63.05

b)

Figure 3: Comparison with Bazzani *et al*. [4].

We started by analysing the performance of the model while varying α which is the free parameter of the model and is directly linked to the probability of creating new groups. In order to find an appropriate value, a performance analysis has been made varying α between 1 and 20. The curves for 1-FP, 1-FN and GDSR obtained varying α are shown in Figure 3(a). The performed analysis highlighted that the GDSR shows a peak at $\alpha = 2$ and in correspondence to that value both 1-FP and 1-FN compare favorably to the results obtained by [4]. As a result, the free parameter α has been set to 2 for all the experiments presented in this section. The results of the comparison with [4] on the `eth` sequence are reported in Figure 3(b). The proposed method outperforms the benchmark on the `eth` sequence for all the considered metrics, proving the efficacy of the proposed approach in detecting groups of different size.

Figure 4 shows the qualitative results on `eth` (first row), `hotel` (second row), and `zara01` (last row). First of all, it is worth noticing that groups (in orange) are correctly estimated when compared with the ground truth (in green), even in the most challenging cases where groups are very close to each other (for example, frame 5148 of `eth` and frame 1753 of `zara01`). The method is also able to deal with a varying number of groups in the scene (from 0 to 4 groups in the figure) and with multiple individuals walking alone (frame 7868 of `eth`). An example of false negative is reported in frame 2876 of `hotel`. As can be seen by later frames (from 2904 to 3424) the model corrects the false negative by detecting the group. This happens because a new component is associated to the group and gains more importance over time being supported by the observed data for several subsequent frames. An example of false positive is reported in frame 41 of `zara01`. Even in this case, the model corrects the error by exploiting both the evidence coming from later frames and employing the social constraint.

All the sequences have been processed with a Matlab implementation on a Xeon E5620 2.4 GHz with 12 GB RAM. Results are produced in real-time up to 42 fps when starting from the output of a people detection algorithm.

Discussion. In this paper, we presented an approach aimed at detecting group formations in video sequences, where the pedestrian have been detected beforehand. The approach is based on Dirichlet Process Mixtures, whose inference has been sped up through a sequential variational scheme. The approach compares favourably with the nowadays state-of-the-art approaches, setting the best performance on some datasets. It is worth noting that our method works online and with real-time performance, while all the other approaches are based on batch processing. This promotes its usage in many real-world commercial application scenarios, such as video surveillance.

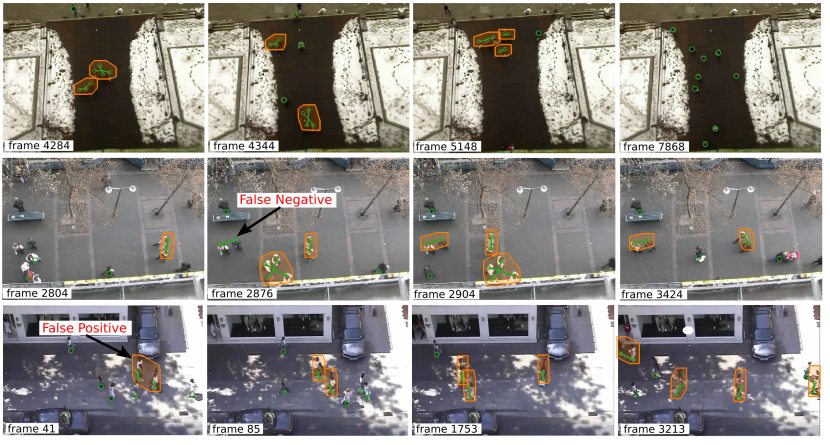


Figure 4: Qualitative results on `eth` (first row), `hotel` (second row), and `zara01` (last row). The ground truth position of individuals and groups are shown with green circles and segments, respectively. The estimated groups are depicted as orange convex hulls (best viewed in colors).

References

[1] A. Ahmed and E.P. Xing. Dynamic non-parametric mixture models and the recurrent chinese restaurant process with applications to evolutionary clustering. In *Proceedings of the 8th SIAM International Conference on Data Mining*, 2008.

[2] C.E. Antoniak. Mixtures of Dirichlet Processes with Applicatioins to Bayesian Non-parametric Problems. *The Annals of Statistics*, 2(6):1152–1174, 1974.

[3] L. Bazzani, M. Cristani, G. Pagetti, D. Tosato, G. Menegaz, and V. Murino. Analyzing groups: a social signaling perspective. In *Video Analytics for Business Intelligence*, Studies in Computational Intelligence. Springer-Verlag, 2012.

[4] L. Bazzani, V. Murino, and M. Cristani. Decentralized particle filter for joint individual-group tracking. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2012.

[5] D.M. Blei and M.I. Jordan. Variational Inference for Dirichlet Process Mixtures. *Bayesian Analysis*, 1(1):121–144, 2006.

[6] A. Braun, S.R. Musse, L.P.L. de Oliveira, and B.E.J. Bodmann. Modeling individual behaviors in crowd simulation. In *Computer Animation and Social Agents, 2003. 16th International Conference on*, pages 143 – 148, may 2003.

[7] W. G. Choi, K. Shahid, and S. Savarese. What are they doing?: Collective activity classification using spatio-temporal relationship among people. pages 1282–1289, 2009.

[8] M. Cristani, L. Bazzani, G. Pagetti, A. Fossati, D. Tosato, A. Del Bue, G. Menegaz, and V. Murino. Social interaction discovery by statistical analysis of f-formations. In *British Machine Vision Conference (BMVC)*, 2011.

[9] T.S. Ferguson. A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1(2):209–230, March 1973.

[10] W. Ge, R.T. Collins, and R.B. Ruback. Vision-based analysis of small groups in pedestrian crowds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 99 (PrePrints), 2011. ISSN 0162-8828.

[11] E.T. Hall. *The Hidden Dimension*. 1966.

[12] D. Helbing and P. Molnar. Social force model for pedestrian dynamics. *PHYSICAL REVIEW E*, 51:4282, 1995.

[13] J.C.S. Jacques, A. Braun, J. Soldera, S.R. Musse, and C.R. Jung. Understanding people motion in video sequences using voronoi diagrams: Detecting and classifying groups. *Pattern Analysis Applications*, 10(4):321–332, 2007.

[14] I. Karamouzas and M. Overmars. Simulating and evaluating the local behavior of small pedestrian groups. *Visualization and Computer Graphics, IEEE Transactions on*, 18(3): 394 –406, march 2012. doi: 10.1109/TVCG.2011.133.

[15] A. Kendon. *Conducting Interaction: Patterns of behavior in focused encounters*. 1990.

[16] A. Lerner, Y. Chrysanthou, and D. Lischinski. Crowds by example. volume 26, pages 655–664, Sep 2007.

[17] J.M. Levine and R.L. Moreland. Small groups. In D.T. Gilbert, S.T. Fiske, and G. Landzey, editors, *The Handbook of Social Psychology*, volume II. McGraw-Hill, 4th edition, 1998.

[18] P. Liang and D. Klein. Online EM for unsupervised models. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 611–619, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[19] R.M. Neal and G.E. Hinton. A new view of the EM algorithm that justifies incremental and other variants. In Michael I. Jordan, editor, *Learning in Graphical Models*. MIT Press, 1998.

[20] S. Pellegrini, A. Ess, K. Schindler, and L. van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *International Conference on Computer Vision*, 2009.

[21] S. Pellegrini, A. Ess, and L. Van Gool. Improving data association by joint modeling of pedestrian trajectories and groupings. In *Proceedings of the 11th European conference on Computer vision: Part I*, ECCV'10, pages 452–465, Berlin, Heidelberg, 2010. Springer-Verlag.

[22] W.D. Penny. Variational Bayes for d-dimensional Gaussian Mixture Models. Technical report, Wellcome Department of Cognitive Neurology - University College London, July 2001.

[23] C.W. Reynolds. Flocks, herds and schools: A distributed behavioral model. *SIGGRAPH Comput. Graph.*, 21(4):25–34, August 1987.

[24] J. Sethuraman. A Constructive Definition of Dirichlet Priors. *Statistica Sinica*, 4:639–650, 1994.

[25] J. Sochman and D.C. Hogg. Who knows who - inverting the social force model for finding groups. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 830 –837, nov. 2011.

[26] A. Vinciarelli, M. Pantic, and H. Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing Journal*, 27(12):1743–1759, 2009.

[27] G. Wang, A. Gallagher, J. Luo, and D. Forsyth. Seeing people in social context: Recognizing people and social relationships. In *European Conference on Computer Vision*, pages 169–182, 2010.

[28] K. Yamaguchi, A.C. Berg, L.E. Ortiz, and T.L. Berg. Who are you with and where are you going? In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2011.